HEREDITY VOLUME 29 PART 2 OCTOBER 1972

VARIATION IN WILD POPULATIONS OF *PAPAVER DUBIUM*
V. THE APPLICATION OF FACTOR ANALYSIS TO THE STUDY OF VARIATION

J. S. GALE and L. J. EAVES
Department of Genetics, University of Birmingham, England

Received 28.x.71

1. Introduction

In a previous investigation (Gale and Arthur, 1972), an attempt was made to estimate genetic correlations between ten characters, taken in pairs, in *Papaver dubium*. Plants were raised from seed collected from six natural populations. From every population, a number of pairs of plants were chosen at random and plants comprising a pair crossed to one another to give a single family. In all, 18 families were raised, a family consisting of four plants. In order to minimise the effects of environmentally induced variation and covariation, correlations were calculated on family means. From these correlations, it was apparent that five of the characters fell into three independent groups. The remaining characters showed significant intermediate sized correlations with the members of more than one group. It was concluded that there were at least three independent sets of loci controlling these characters. Characters correlated with members of more than one group are presumably controlled by loci from more than one set. We may note that this procedure depends on the assumption that the correlation coefficient gives an adequate summary of the relationship between any two of our characters; we shall discuss this point in more detail below.

Now nine of the ten characters showed significant genetical variation within populations, these nine falling into the three separate groups described above. It follows that three different polymorphisms have been detected in these populations. However, the number of families per population was small, so that it was not possible to decide how many of these polymorphisms existed in a given population. Accordingly, it was decided to make a more detailed study of a single population, with a view to making a minimum estimate of the number of independent polymorphisms in that population. The general procedure was the same as in the earlier experiment, but incorporated two improvements.

Firstly, the number of plants per family was increased to 20. This large family size seems essential, at least in the case of poppies, if really accurate results are to be obtained. In detail, the estimated variance of family means for any given character, has expected value

$$\sigma_B^2 + \frac{1}{K}\sigma_w^2$$

where

σ_B^2 = population variance of true family means

σ_w^2 = population average variance within families

K = family size.

Analogous expressions hold for estimated covariances of family means for a pair of characters; the expected value has the same form but with population

variances replaced by covariances. Hence, if the family size is sufficiently large, the contribution of within family variation and covariation, which is partly environmental, to our correlation coefficient, calculated on family means, will be negligible. Our coefficient is thus an estimate of correlation which is purely genetic in origin; we may refer to this as a genetic correlation coefficient, although this term is more strictly used for the correlation between breeding values (see *e.g.* Falconer, 1960).

As an alternative to the use of large families, we might attempt to estimate within family variances and covariances. These estimates are then divided by the family size and subtracted from the appropriate variances and covariances of family means. Correlations are then calculated from these adjusted estimates (Mode and Robinson, 1959). The efficacy of this procedure will depend to some extent on the relative sizes of between and within family components. In the case of poppies, the procedure fails if family sizes are small, since the estimates have very large sampling errors. This may be shown as follows. While the true correlations cannot, of course, exceed unity in absolute value, no such restraint applies to their estimates since these are not obtained, in this procedure, directly from paired comparisons and are not therefore subject to the restrictions imposed by Cauchy's inequality. In a sense, this is fortunate, since we can obtain an idea of sampling error from the frequency of occurrence of " impossible " estimates. In practice, estimates as large as three in absolute value are quite common (Ooi and Arthur, personal communication).

Given then that we have decided to raise large families, the adjustments to variances and covariances just described may still be made, although there are disadvantages in doing this. As far as we are aware, little is known of the probability distribution of the resulting estimates of correlation, apart from their variances. Thus the estimates do not lead to any tests of significance. Accordingly, it seems best to raise families sufficiently large for estimates of correlation calculated on unadjusted estimates to differ very little from those calculated from adjusted estimates. The adjustments are then unnecessary and we thus have the usual estimated correlation coefficient with well-known properties. With families of size 30, the difference between the two types of estimate turned out to be trifling, for all five characters measured in 1970 (Bassi, personal communication). As we shall show, for families of size 20, the difference is still trivial.

A second improvement was in the procedure used to allocate characters to groups. While characters showing near-zero genetic correlation *inter se* are readily allocated to different groups, characters under the control of genes belonging to more than one set (*i.e.* characters belonging to two or more groups) are difficult to assign. Further, it is difficult to decide, in the case of a character showing genetic correlation with other characters, whether there is any variation present specific to that character or whether *all* variation in that character is necessarily associated with variation in others. For these reasons, we have used the technique of factor analysis in order to obtain a more objective picture of the relationship between our characters.

2. Materials and methods

The plants used were derived from an experimental population set up for another purpose in 1964. A previously grassed-over plot, measuring

30 × 4 feet was cleared. About 100,000 seeds from 100 plants growing on the University campus at a distance of 300-400 yards from the plot were scattered on the plot (Lawrence, personal communication). Since that time, the population was maintained by self-seeding and weeding was kept to the absolute minimum necessary to prevent poppies from being crowded out. Thus we are dealing here with a derivative of the University campus population mentioned in the earlier paper, maintained in close proximity to the parent population.

In 1968, it was decided to set up a number of inbred lines, starting with plants growing on the plot. Accordingly, 20 randomly chosen plants were selfed and progeny grown on the experimental field in 1969. The lines were maintained by further selfings in 1969 and 1970, thus giving rise to 20 families, each derived from a single wild plant. These families were raised, on the experimental field, in 1971, 20 plants being grown per family, thus giving 400 plants in all, these being the material of the present study.

It was intended, as far as possible, to measure the same ten characters as on the previous occasion. However, it proved impossible to measure three of these, namely juvenile elevation, leaf number at flowering time and diameter at flowering time, under field conditions. For example, the number of leaves at flowering time exceeded 300 in some plants and it became apparent that it would be necessary to pull the plant to pieces in order to obtain an accurate count.

In the earlier study, juvenile characters were measured at both 7 weeks and 8 weeks after sowing. However, since performance at these two times was very similar, we have confined ourselves to measurements at 7 weeks only. These were made immediately before planting. On the other hand, results obtained in 1967 suggested that the pattern of correlations obtained at 10 weeks, or so, was rather different from that at 7 weeks (Arthur, personal communication). To some extent this might be expected, at least for correlations involving juvenile height. For the height at about 7 weeks in fact represents the height of the outermost leaves, which gradually become less elevated as the plant develops further and lie almost flat at about 10 weeks. At this latter period, the highest leaves are those in the crown, which is beginning to elongate at this time. Since at 10 weeks this elongation is not very advanced in some plants, we have the paradoxical result that, in some cases, height at 10 weeks is considerably less than height 3 weeks earlier. We concluded, therefore, that measurements at both 7 and 10 weeks were desirable. These measurements, of leaf number, height and diameter, will be denoted LN7, H7, D7 and LN10, H10, D10 for the two occasions of measurement respectively.

We now turn to measurements made at flowering time. Here we were faced with a tiresome problem, in that, on many plants, first and sometimes second flowers had been removed, in the bud stage, by birds. However, the third flower is easily recognised as such and we decided, therefore, to score flowering time (FT) as the time of opening of the *third* flower. Humphreys (personal communication) has found, in plants raised from seed collected from two natural populations, that time of opening of the first and third flowers show a correlation of over 0·9 *inter se* and that the correlations of time of opening of third flower with other characters are closely similar to the corresponding correlations obtained for the first flower. Thus our results should be comparable with earlier results obtained by scoring the first

flower. Other characters scored at flowering time were plant height (HF), anther number in the third flower (AF) and number of buds (BF).

Finally, we also scored two characters first studied by Lawrence (1972), namely number of stigmatic rays (SR), which in our case was the total number of rays on the gynaecia of the third, fourth and fifth flowers to open, and also capsule number (CN).

3. Estimated correlation coefficients

Estimated between family ($\hat{\sigma}^2_B$) and within family ($\hat{\sigma}^2_w$) components of variation are given in table 1. All between family components are significant at the 0·1 per cent. level. For the purposes of comparison, we include estimates of σ^2_B obtained in the earlier study.

Table 1

Components of variation

Character	$\hat{\sigma}^2_w$	$\hat{\sigma}^2_B$	$\hat{\sigma}^2_B$ (previous study)
LN7	1·8	1·4	1·5
H7	153·2	38·3	29·1
D7	442·3	106·7	188·7
LN10	16·3	18·1	—
H10	153·2	355·2	—
D10	933·4	812·4	—
FT	6·7	13·9	62·6
HF	3604·9	2787·4	2330·0
AF	159·1	131·7	115·0
BF	51·0	42·1	0·0
SR	1·7	0·5	—
CN	684·1	129·2	—

$\hat{\sigma}^2_B$ = between family component.
$\hat{\sigma}^2_w$ = within family component.

In view of the large sampling errors associated with these estimates, the agreement between values for $\hat{\sigma}^2_B$ for the two experiments are surprisingly good, particularly since measurements were made under very different conditions. Certainly, discrepancies as large as this have been found when seed from a group of families was raised in two blocks sown 1 week apart, every family appearing in both blocks, and comparisons were made between estimated components calculated for the two blocks separately (Gale and Arthur, 1972). Tentatively, then, we may conclude that, for most characters, the genetic variance in our population is similar to the average genetic variance for our previous six populations taken together. In this sense, our population is a " typical " one. On the other hand, the low variance for flowering time in the population under study probably represents a real departure from the situation in the other five populations, since we have other evidence suggesting this. The other obvious discrepancy, in number of buds at flowering time, may represent a real difference between populations or, alternatively, either some form of genotype-environmental interaction or simply a reflection of the different methods by which flowering time was scored in the different experiments.

We turn now to the estimated correlations between characters. In table 2, we give the " raw " genetic correlations, that is, the correlations

calculated directly on family means. In table 3 are given the " adjusted " genetic correlations, that is correlations adjusted to allow for within family variances and covariances, as discussed in section 1.

It will be seen that the adjusted estimates differ little from the raw estimates in any individual case, although in general adjusted estimates are slightly larger in absolute value than the raw estimates. This happens

TABLE 2

Correlations calculated on family means

	LN7	H7	D7	LN10	H10	D10	FT	HF	AF	BF	SR	CN
LN7		−0·07	0·01	0·22	−0·18	−0·12	0·32	−0·22	−0·10	0·68	−0·26	0·53
H7			0·60	0·37	0·63	0·62	−0·52	−0·37	−0·16	−0·46	−0·12	0·13
D7				0·34	0·60	0·57	−0·52	−0·25	−0·05	−0·26	−0·02	−0·03
LN10					0·75	0·76	−0·61	−0·59	−0·54	−0·13	−0·19	0·28
H10						0·90	−0·78	−0·49	−0·43	−0·44	−0·07	0·01
D10							−0·84	−0·58	−0·34	−0·43	−0·03	0·19
FT								0·73	0·27	0·70	0·02	0·00
HF									0·38	0·42	0·34	−0·46
AF										0·13	0·71	−0·18
BF											0·02	0·18
SR												−0·40
CN												

because within family correlations are usually low. It does not necessarily follow from this that environmental correlations are low, since the within family correlations will be depressed by the effects of errors of measurement, which should, in general, be uncorrelated.

When scatter diagrams for family means of characters taken in pairs were plotted, it became apparent that the relationship, if any, between a pair of

TABLE 3

Correlations adjusted for within family effects

	LN7	H7	D7	LN10	H10	D10	FT	HF	AF	BF	SR	CN
LN7		−0·08	−0·06	0·22	−0·18	−0·14	0·34	−0·22	−0·11	0·72	−0·28	0·60
H7			0·70	0·41	0·68	0·68	−0·58	−0·43	−0·19	−0·52	−0·15	0·14
D7				0·36	0·66	0·62	−0·57	−0·28	−0·05	−0·30	−0·02	−0·05
LN10					0·77	0·78	−0·62	−0·62	−0·57	−0·14	−0·20	0·32
H10						0·93	−0·80	−0·51	−0·45	−0·46	−0·08	0·02
D10							−0·87	−0·62	−0·37	−0·46	−0·04	0·20
FT								0·75	0·28	0·71	0·02	0·01
HF									0·39	0·42	0·36	−0·54
AF										0·13	0·76	−0·21
BF											0·02	0·18
SR												−0·49
CN												

characters was essentially linear. That is, although occasionally a line showing a small degree of curvilinearity would give a slightly better fit than a straight line, no cases were found where the line was U-shaped. Thus, in the present case, the genetic correlation coefficient gives a reliable indication of the degree of relationship between genotypic values for a pair of characters.

However, it does not follow from this that if two characters, say X and Y, are under the control of exactly the same set of genes, the genetic correlation between them will be unity, even in the absence of sampling error. Indeed, in principle, the correlation could be zero (Falconer, 1960). Consider, for

example, the following very special case. Suppose both characters are controlled by genes at n loci, with two alleles per locus and that allele frequencies at each locus are $\frac{1}{2}$. Suppose further that at half the loci, the increasing allele for X is also the increasing allele for Y, whereas, at the other loci, the increasing allele for X is the decreasing allele for Y. If the effect of any gene substitution is to alter X or Y, either upwards or downwards, by one unit and gene action is additive, a zero correlation will result.

This special case illustrates a general point, namely that if two characters are controlled by the same set of genes, a zero correlation will result only if there is a nice adjustment between effects of gene substitutions and gene frequencies. This is not very helpful, however, because even if the adjustment is not very precise, a lowish correlation could still be found in some cases. Thus, if we wish to interpret a low correlation between X and Y as implying that these two characters are, in the main, under the control of separate sets of loci, we shall have to make some assumption about the loci concerned. We shall make the following " postulate of consistency of gene action " for the case where X and Y are under the control of the same set of genes. We postulate that, at most loci having a substantial effect on the two characters either (1) increasing (decreasing) alleles for X are consistently increasing (decreasing) alleles for Y at the loci under consideration or (2) increasing (decreasing) alleles for X are consistently decreasing (increasing) alleles for Y at the loci under consideration.

However, even given this postulate, the correlation could still be some way from unity. For whereas we have supposed a consistency over loci with regard to the *direction* of gene effects, it would be quite wrong to assume further a consistency in *magnitude* of effect. If, then, we plot a scatter diagram of genotypic values of X and Y, the points will not usually be on a straight line but will give the usual elliptical pattern, even in the absence of sampling error, giving rise to a correlation of intermediate size.

To sum up, a high correlation indicates pleiotropic effects of one set of loci or perhaps linkage between two independent sets. A very low correlation indicates two separate sets of loci controlling the characters at some stage, provided the consistency postulate is applicable. An intermediate correlation is difficult to interpret; it may indicate the pleiotropic effect of one set of loci or linkage between two sets of loci or one set of loci controlling both characters in conjunction with two other independent sets, one for each character.

We shall assume that the consistency postulate is appropriate but will make no other assumptions. We are thus following the approach tacitly assumed in the earlier paper; if the correlation between a pair of characters is very low, we shall conclude that these characters are controlled by two distinct sets of loci. If a pair of characters show an intermediate or high correlation *inter se*, we shall regard them as controlled by the same set of loci. Thus we shall obtain a *minimum* estimate of the number of sets of loci controlling our characters, that is, a minimum estimate of the number of polymorphisms present in our original population. Our estimate may, of course, be reduced by the effects of linkage between loci which in fact belong to distinct sets.

As explained above, we shall attempt to separate our characters into groups, each group representing a distinct set of loci, by means of factor analysis. As this procedure has not been widely used in genetical problems it may be helpful to give an account of the method and of the computations

required. A helpful account of basic notions is given in Morrison (1967). A comprehensive survey of the field, and in particular of recent developments, will be found in Lawley and Maxwell (1971).

4. Factor analysis: procedure

Suppose we have investigated the variation in p characters. In most experiments, a " character " will be a measurement that varies from one individual to another. In the present context, however, a " character " is the family mean for a given measurement, this mean varying from one family to another. In general, the variation in a given character, say X, will be accompanied, to some extent, by correlated variation in one or more other characters. In that the true correlations between X and these other characters will normally be less than unity in absolute value, we may regard the variation in X as made up of two parts, namely the variation which is associated with variation in other characters (" common variation ") and variation which is independent of variation in other characters (" specific variation "). Let us suppose, then, that the specific variation has been calculated for every character. We are left with all variation which is common to two or more characters. Generally, not all characters will show common variation with each other; for example, a character X may show common variation with a character Y but not with a character Z. The latter, however, may show common variation with some other characters, say A and B. Thus we are led to the idea of accounting for all the common variation in terms of variation in k underlying uncorrelated factors. Let X_1, X_2, ..., X_p be the values of the various characters in some " individual "; in the present case, an " individual " is an individual family. The X's vary from " individual " to "individual", thus giving rise to the usual mean and standard deviation of any X. Similarly, we regard the factors as varying in the same way; factors are defined so that each has variance unity. Finally, we have the specific variation for every character, measured by the " specific variance " for that character; in many cases, this turns out to be zero, as we shall discuss later.

Let f_r be the rth factor. We postulate the linear model

$$\begin{aligned} X_1 &= \lambda_{11}f_1+\lambda_{12}f_2+\ldots+\lambda_{1k}f_k+e_1 \\ &\vdots \\ X_p &= \lambda_{p1}f_1+\lambda_{p2}f_2+\ldots+\lambda_{pk}f_k+e_p. \end{aligned} \tag{1}$$

Here the λ's are unknown constants and the e's represent specific variation *e.g.* e_1 represents the difference between the actual value of X_1 and the value predicted from the factors. The specific variance for the ith character is defined as

$$\psi_i = \text{var}\,(e_i). \tag{2}$$

If the X's are standardised (*i.e.* each is measured from its mean and divided by its standard deviation), then the λ's become correlation coefficients between characters and factors, *e.g.* λ_{sm} is the correlation between the sth character and the mth factor; it is called the loading of this character on this factor.

In matrix notation, we may write (1) as

$$\mathbf{X} = \Lambda\mathbf{f}+\mathbf{e}. \tag{3}$$

Then it follows that the covariance matrix of the X's is

$$\Sigma = \Lambda\Lambda' + \Psi \tag{4}$$

where Ψ is a diagonal matrix, with diagonal elements $\psi_1, \psi_2, \ldots, \psi_p$. If the X's are standardised, Σ is the correlation matrix of the X's. If, as is often the case, different characters are measured in different units, their correlation matrix is the natural starting point for the analysis. Fortunately, most of the calculations are the same, whether we start with covariances or with correlations. We shall, for simplicity, suppose that we start with the correlation matrix.

We should like to know (1) the number of factors (2) the specific variances and (3) the loadings. Now it turns out that if we postulate that there are k factors, we can estimate the elements of $\Lambda\Lambda'$ and of Ψ by maximum likelihood. Thus we can obtain predicted values for all the elements of Σ. Of course Σ is unknown; we actually have $\mathbf{S}$ the observed correlation matrix. We therefore test whether the elements of $\mathbf{S}$, taken together, differ significantly from the corresponding elements of the *predicted* Σ. If there is a significant difference, our value of k is incorrect. The procedure, therefore, is to start with $k = 1$ and test the departure of $\mathbf{S}$ from the predicted Σ; if this is significant, we take $k = 2$ and repeat the procedure. We continue in this way, augmenting k by unity, until such time as the difference between $\mathbf{S}$ and predicted Σ is not significant. If, however, we have some prior reason (*e.g.* previous experience with the characters concerned) for expecting k to take some particular value, it would usually be best to start with that value.

Thus the number of factors and the specific variances can be estimated. However, the loadings cannot be found unless we make further assumptions. Let $\mathbf{T}$ by any orthogonal matrix, that is, a matrix such that $\mathbf{TT}'$ is the unit matrix. Consider a set of loadings Λ and multiply Λ by $\mathbf{T}$ to obtain new loadings $\Lambda\mathbf{T}$. These have the property

$$(\Lambda\mathbf{T})(\Lambda\mathbf{T})' = \Lambda\mathbf{T}\mathbf{T}'\Lambda' = \Lambda\Lambda'. \tag{5}$$

Thus the new loadings give the same $\Lambda\Lambda'$ as the old and therefore predict Σ just as well (or badly). Our choice of loadings is, therefore, in the first instance, arbitrary; any will do, provided they give the right $\Lambda\Lambda'$, which is all we require for estimating number of factors and the specific variances. Now it may be shown that among all the appropriate Λ, there must be a Λ satisfying

$$\Lambda'\Psi^{-1}\Lambda = \text{a diagonal matrix.} \tag{6}$$

This condition was introduced by Lawley (1940). In practice, some condition is required, since we have to estimate the Λ in order to estimate the $\Lambda\Lambda'$. It should be emphasised that this condition has no biological meaning and attempts to interpret the Λ obtained subject to this condition are usually pointless (but see below).

We turn now to the actual process of maximum likelihood estimation. We should emphasise that this must be done on the *raw* correlations. Exaggerated rounding off of the latter leads to difficulties; it is best, in the case of correlation matrices, not to round off the original values to less than three decimal places. We have, in fact, retained four places in the computations to be described later.

Let L be the logarithm of the likelihood, apart from an additive constant. We wish to maximise L, for variations in the $\boldsymbol{\Psi}$ and in the $\boldsymbol{\Lambda}$, subject, of course, to (6). Equations which the maximum likelihood estimates must satisfy were first given by Lawley (1940). However, it became apparent that these equations did not lead to a satisfactory procedure for calculating the estimates. Recently, however, new procedures have been proposed (Jöreskog, 1967; Jöreskog and Lawley, 1968; Clarke, 1970), which are very satisfactory.

It proves convenient to minimise

$$F(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = -\frac{2}{n} L - \log |\mathbf{S}| - p \tag{7}$$

where n is the number of " individuals " observed, minus one; as before $\mathbf{S}$ is the observed correlation (or covariance) matrix with determinant $|\mathbf{S}|$ and p is the number of characters. Since n, $|\mathbf{S}|$ and p are all constants, the values of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ which maximise L will also minimise $F(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$.

Suppose first that the $\boldsymbol{\Psi}$ are given. Consider the matrix

$$\boldsymbol{\Psi}^{-\frac{1}{2}}\mathbf{S}\boldsymbol{\Psi}^{-\frac{1}{2}}$$

with latent roots, in *descending order* of magnitude

$$\theta_1, \theta_2, \ldots, \theta_p$$

and corresponding latent column vectors

$$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p.$$

Let w_{ij} be the element in the ith row of $\mathbf{w}_j$ and let the vectors be standardised so that, for any j,

$$\sum_{i=1}^{p} w_{ij}^2 = 1. \tag{8}$$

Then (Jöreskog, 1967) the maximum likelihood estimates of the $\boldsymbol{\Lambda}$, conditional on the given $\boldsymbol{\Psi}$, are readily found. To do this, we consider the k *largest* latent roots and construct a matrix $\boldsymbol{\Theta}$ with $\theta_1, \theta_2, \ldots, \theta_k$ on the leading diagonal and zero elsewhere. We further construct a matrix $\boldsymbol{\Omega}$, of which the first column is $\mathbf{w}_1$, the second column $\mathbf{w}_2$ and so on, the last column being $\mathbf{w}_k$. Jöreskog then shows that $\hat{\boldsymbol{\Lambda}}$, the estimates of the $\boldsymbol{\Lambda}$ conditional on the given $\boldsymbol{\Psi}$, are given by

$$\hat{\boldsymbol{\Lambda}} = \boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\Omega}(\boldsymbol{\Theta} - \mathbf{I})^{\frac{1}{2}} \tag{9}$$

where $\mathbf{I}$ is the unit matrix.

It turns out that, if we substitute the given $\boldsymbol{\Psi}$ and our conditional estimates $\hat{\boldsymbol{\Lambda}}$ into $F(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$, as given by (7), we obtain

$$f(\boldsymbol{\Psi}) = \sum_{j=k+1}^{p} [\theta_j - \log \theta_j - 1] \tag{10}$$

this being the minimum value of $F(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$, conditional on the given $\boldsymbol{\Psi}$.

In order then to obtain maximum likelihood estimates of the $\boldsymbol{\Psi}$ we must minimise $f(\boldsymbol{\Psi})$ for variations in the $\boldsymbol{\Psi}$. Once these maximum likelihood estimates of the $\boldsymbol{\Psi}$ have been obtained, the true maximum likelihood estimates of the $\boldsymbol{\Lambda}$ can be found from (9). The values of $\boldsymbol{\Psi}$ which minimise

$f(\boldsymbol{\Psi})$ can be found numerically using the Newton-Raphson method, for which appropriate formulae have been developed by Clarke (1970). This method is probably the most rapid available, in cases where the number of characters is not too large (Lawley and Maxwell, 1971). The Newton-Raphson method, when used in maximum likelihood estimation, is known in the genetical literature as Fisher's scoring method (see *e.g.* Mather, 1951; Bailey, 1961) and it may be helpful to use Fisher's terminology here, provided it is understood that by " score " we mean maximum likelihood score (and not factor score).

For convenience, we shall write f as short for $f(\boldsymbol{\Psi})$. Then it may be shown (Jöreskog, 1967; Clarke, 1970) that

$$\frac{\partial f}{\partial \psi_i} = \frac{1}{\psi_i} \sum_{j=k+1}^{p} (1-\theta_j) w_{ij}^2. \tag{11}$$

With the same notation as before, let

$$\boldsymbol{\Phi} = \boldsymbol{\Psi}^{-\frac{1}{2}}(\mathbf{I} - \boldsymbol{\Omega}\boldsymbol{\Omega}')\boldsymbol{\Psi}^{-\frac{1}{2}}. \tag{12}$$

If ϕ_{il} is the element in the ith row, lth column of $\boldsymbol{\Phi}$ it may be shown (Clarke, 1970 and personal communication) that

$$\beta_{il} = \frac{\partial^2 f}{\partial \psi_i \partial \psi_l} = \phi_{il}^2 - \frac{2\delta_{il}}{\psi_i}\frac{\partial f}{\partial \psi_i} + \frac{1}{\psi_i \psi_l}\left[\sum_{j=1}^{k} w_{ij}w_{lj} \sum_{r=k+1}^{p} \frac{2\theta_j(\theta_r - 1)}{\theta_r - \theta_j} w_{ir}w_{lr}\right] \tag{13}$$

where $\delta_{il} = 1$ if $i = l$ and zero otherwise.

Now let $\hat{\boldsymbol{\Psi}}_{(0)}$ be a vector of trial values for the maximum likelihood estimates of the $\boldsymbol{\Psi}$. Appropriate values are given by

$$\hat{\Psi}_i = \frac{1 - k/(2p)}{s^{ii}} \tag{14}$$

(Jöreskog, 1963), where s^{ii} is the ith diagonal element in $\mathbf{S}^{-1}$. Let $\partial f/\partial \psi_i$, ϕ_{il}^2 and β_{il} be evaluated using these trial values. Denote the column vector with $\partial f/\partial \psi_i$ in its ith row as $\mathbf{T}$, the matrix with ϕ_{il}^2 in its ith row, lth column as $\mathbf{G}$ and the matrix with β_{il} in its ith row, lth column as $\mathbf{J}$, all elements being evaluated at the trial values given by (14). Then in Fisher's terminology, the matrix of scores is

$$-\frac{n}{2}\mathbf{T}$$

and the (approximate) matrix of information *realised* is

$$\frac{n}{2}\mathbf{J}$$

whence an improved set of estimates, say $\hat{\boldsymbol{\Psi}}_{(1)}$, is given, in the usual way by

$$\hat{\boldsymbol{\Psi}}_{(1)} = \hat{\boldsymbol{\Psi}}_{(0)} - \mathbf{J}^{-1}\mathbf{T} \tag{15}$$

with one important proviso, that $\mathbf{J}$, as evaluated, shall be a positive definite matrix (if so, all its latent roots will be positive). Whereas if $\mathbf{J}$ is evaluated using the true maximum likelihood estimates of the $\boldsymbol{\Psi}$, it must be positive

definite, this will not necessarily hold for trial estimates some distance from the true estimates. However, this difficulty is easily overcome (Clarke, 1970); if $\mathbf{J}$ is not positive definite, $\mathbf{G}^{-1}$ is substituted for $\mathbf{J}^{-1}$ in (15), since $\mathbf{G}$ is always positive definite.

Once improved estimates of the $\boldsymbol{\Psi}$ have been obtained, they may be used as new trial values and the whole process repeated to give still better estimates. By continually repeating the process, we may obtain estimates to any desired degree of accuracy. Clarke suggests using $\mathbf{G}^{-1}$ rather than $\mathbf{J}^{-1}$ for the first iteration and also after an iteration in the course of which the estimate of any ψ_i changes by more than 0·1 in absolute value. This helps to speed up the whole process.

In order to test the agreement between observed and expected, we calculate

$$[n-(2p+5)/6-2k/3]\log_e \frac{|\hat{\boldsymbol{\Psi}}+\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'|}{|\mathbf{S}|} \tag{16}$$

where $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$ are our maximum likelihood estimates. If n is large, this will be approximately a χ^2 for

$$\tfrac{1}{2}[(p-k)^2-p-k]$$

degrees of freedom. Lawley and Maxwell (1971) state that the approximation is probably good enough if $(n-p) \geqq 50$. Thus, in our present case where $n = 19$, the use of this approximation is, in principle, quite unsatisfactory. Fortunately, however, this turns out not to be a serious difficulty with our data (see below).

An old difficulty in factor analysis is that, in order to maximise the likelihood, it may be necessary for some of the $\hat{\psi}_i$ to be negative. Since the ψ_i are variances, negative estimates are inadmissable. In such cases, the best estimate of these ψ_i is zero. The appropriate procedure is to take these ψ_i to be zero from the start and carry out the analysis on this basis. The method for doing this is given by Jöreskog (1967). The first step is to identify the characters which have zero specific variances. If, at any stage during the standard analysis described above, a specific variance becomes negative it is set " on the boundary " *i.e.* at some small positive value, say 0·001, and kept at that value until the standard analysis is completed. If there are no such cases, the analysis is concluded. Otherwise, we have an " improper solution ", in which a number of variables, say m, have specifics on the boundary. It is these specifics which are now set equal to zero from the start of our re-analysis.

It is easier if we list the variables so that the m variables are written first. Correlations involving these m, but not other characters, appear as $\mathbf{S}_{11}$, a submatrix of $\mathbf{S}$. Similarly, correlations involving only the other $(p-m)$ characters appear as a submatrix $\mathbf{S}_{22}$. Since we have listed the m variables first, we may write $\mathbf{S}$ in the form

$$\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}$$

where $\mathbf{S}_{12}$ $(= \mathbf{S}'_{21})$ is a submatrix made up of correlations of the m with the $(p-m)$.

For the $\mathbf{S}_{11}$, we must find factors which account for all the variation (since specifics are zero). This is the situation appropriate for principal

components analysis (see *e.g.* Morrison, 1967). In order to account for *all* the variation represented by $\mathbf{S}_{11}$, we shall require m factors. Let $\mathbf{D}$ be a matrix with the latent roots of $\mathbf{S}_{11}$ on the leading diagonal and zero elsewhere and let $\mathbf{C}$ be the corresponding matrix of column vectors, standardised so that the sum of the squared elements in a column is unity. The loadings of the m characters on the m factors are, from principal components theory, given as

$$\hat{\Lambda}_{11} = \mathbf{S}_{11}\mathbf{C}\mathbf{D}^{-\frac{1}{2}}.$$

The loadings of the other $(p-m)$ characters on our m factors may be shown to be

$$\hat{\Lambda}_{21} = \mathbf{S}_{21}\mathbf{C}\mathbf{D}^{-\frac{1}{2}}.$$

We now ask how many more factors, if any, we must invoke to account for the variation in the $(p-m)$ characters. Since we have already accounted for all variation in the m characters, the loadings of these m on the additional factors, if they exist, must be zero. In that any variation which the $(p-m)$ share with the m has been taken care of in the $\hat{\Lambda}_{21}$, we consider only variation in the $(p-m)$ which might be found when the m are kept fixed; we ask how many factors, if any, are required to explain this residual variation. The correlation matrix for the $(p-m)$, given fixed m, is from standard correlation theory

$$\mathbf{S}_{22.1} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}.$$

The m characters are said to have been " partialled out ". We now carry out the standard factor analysis, as if there were only $(p-m)$ variables with correlation matrix $\mathbf{S}_{22.1}$. However, the number of factors we start with will depend on the stage of the analysis. For example, suppose we have already rejected $k = 1$, 2 or 3 and find that m specifics are on the boundary when we try $k = 4$. If $m = 4$, then we should start the analysis of the $(p-m)$ with the number of factors equal to zero. On the other hand, if m were equal to 1, it would be futile to assume less than three factors for the $(p-m)$, since we have already rejected the notion that there are less than four factors.

The procedure described for dealing with specifics on the boundary was first suggested by Lawley. Jöreskog (1967) gives a rigorous demonstration that the approach does give maximum likelihood estimates.

5. Factor analysis of the data

We started our computations by postulating a single factor. No negative specifics were detected. The approximate χ^2 testing goodness of fit of the model turned out to be 98·85 for 54 d.f. with tabulated $P < 0{\cdot}1$ per cent. In spite of reservations about the use of this χ^2 on the present data, P is so small that there can be no serious doubt that the one-factor model must be rejected.

On attempting to fit a two-factor model, we found that three variables gave negative specifics. Hence no proper two-factor solution exists; the presence of three negative specifics means that at least three factors must be postulated.

We therefore started again, this time with three factors. One variable gave a negative specific. This was partialled out, as described in section 4. The χ^2 proved to be 44·25 for 34 d.f. with tabulated P = 20 per cent. –

10 per cent. Thus we have no reason to reject the three-factor model. We conclude that at least three independent polymorphisms exist in this population.

Factor loadings and specific variances are given in table 4. Formulae for the standard errors of these estimates, for cases where n is large are given by Lawley and Maxwell (1971). In the present case, the value of n is too small for the formulae to be appropriate.

TABLE 4

Maximum likelihood estimates of loadings and specific variances

Character	Loadings			Specific variance
	Factor 1	Factor 2	Factor 3	
LN7	0·72	−0·59	−0·22	0·09
H7	−0·46	−0·22	−0·37	0·60
D7	−0·47	−0·34	−0·25	0·60
LN10	−0·30	−0·54	−0·59	0·26
H10	−0·71	−0·39	−0·49	0·11
D10	−0·66	−0·40	−0·58	0·07
FT	0·64	−0·03	0·73	0·05
HF	0·00	0·00	1·00	0·00
AF	0·09	0·23	0·38	0·80
BF	0·61	−0·54	0·42	0·16
SR	−0·25	0·06	0·34	0·82
CN	0·40	−0·29	−0·46	0·55

Even had we been able to establish the significance of some of the specific variances, it would be dangerous to interpret these as evidence for further polymorphisms, each for one character only. If such polymorphisms exist, specific variances will indeed be inflated, but the converse does not hold since specific variances must include effects of non-linear relationships between characters, which are present occasionally to a small degree, and also within family effects which have not been completely eliminated by taking family means.

As we have discussed earlier, although the *number* of factors detected has biological meaning, no biological significance can be attached to the factors themselves or to the corresponding loadings, since these have been estimated using the arbitrary condition (6). If now we drop this condition, we may argue tentatively as follows. Given a group of characters which cover a number of different aspects of the life of a plant, it would be very surprising if most of the characters were to load on all three factors. Rather, we would expect at least some characters to show a high loading on some factor(s) and a low loading on the rest. This pattern of loadings would approximate to the type of pattern known as " simple structure " (Thurstone, 1945).

Since, as discussed earlier, we can multiply the Λ by any orthogonal matrix $\mathbf{T}$ without affecting the goodness of fit of the model, we can attempt to find a $\mathbf{T}$ such that the new loadings $\Lambda\mathbf{T}$ exhibit something approaching simple structure. Geometrically, this is equivalent to representing the factors on mutually orthogonal axes and rigidly rotating these axes; hence multiplying by $\mathbf{T}$ is referred to as " factor rotation ". Details are given in the textbooks cited.

Rotation was carried out by the varimax method of Kaiser (1958) Results are given in table 5. They obviously do not give a neat picture

Factor 1 is the easiest to interpret; all the characters which represent rate of development (H7, D7, LN10, H10, D10, FT) apart, rather oddly, from LN7, load heavily or fairly heavily on it. Factors 2 and 3, however, defy any simple explanation. A careful study of the whole profile of development would probably be the most useful approach to elucidate the whole situation.

Comparing our results with those of Gale and Arthur (1972), we find that although the number of polymorphisms detected is the same in both experiments, the detailed grouping of characters shows differences, sometimes marked. H7, D7, FT and to a lesser extent HF show a fair measure of agreement; the most striking difference is given by LN7, which was clearly associated with H7 but not HF in the previous experiment. Both AF and BF also behave differently. In the previous experiment, AF was closely associated with HF, whereas here, variation in AF is mostly specific. Differences in BF have already been noted and discussed.

TABLE 5

Loadings obtained after factor rotation by the varimax method

	Loadings		
Character	Factor 1	Factor 2	Factor 3
LN7	0·04	−0·68	−0·67
H7	−0·60	0·19	−0·05
D7	−0·63	0·05	0·03
LN10	−0·78	−0·03	−0·36
H10	−0·91	0·22	−0·02
D10	−0·93	0·24	−0·12
FT	0·72	−0·63	0·20
HF	0·41	−0·45	0·80
AF	0·36	−0·03	0·27
BF	0·26	−0·87	−0·10
SR	0·01	0·01	0·43
CN	−0·09	−0·20	0·64

While some of the characters were measured in both experiments, some characters were not, for reasons given in section 2. It has often been noted that a difference of this kind may alter the factor structure for the characters common to both experiments. Some of the changes we have discussed (*e.g.* in AF) seem too drastic for this explanation to be plausible. We conclude that, as is hardly surprising, results obtained from a single population will not, in general, be the same as results obtained by combining material from different populations.

More generally, we should stress that factors and loadings will not necessarily be constant from one situation to another. In so far as different genes may be segregating in different populations and genetic variation between populations may well involve some loci which are usually fixed within populations, different factor structures will emerge. Similarly, results obtained on inbred lines may differ from those on F_1's between them, since correlations between dominance effects may differ from those between additive effects; the derived F_2's and later generations could also be different, owing to recombination. Finally, if characters show strong genotype-environmental interaction, different factor structures would arise if plants were grown in different environments.

6. Summary

1. Plants derived from 20 partially inbred lines derived from a wild population of *Papaver dubium* were scored for 12 characters. All of these characters showed significant differences between lines.

2. Correlation coefficients between characters were calculated on family means. These correlations represent covariation which is almost entirely genetical in origin. Non-linear relationships between characters were virtually absent.

3. On the assumption that gene substitutions affecting a pair of characters are consistent in their action, *i.e.* substitutions which increase one character nearly always increase the other, or nearly always decrease the other, a low genetic correlation between two characters would imply that they are controlled by separate sets of genes. Hence it should be possible to obtain a minimum estimate of the number of sets of loci controlling the characters studied.

4. This estimate is best obtained by means of factor analysis. This method shows that three independent factors must be invoked to account for the observed results.

5. It is concluded that at least three independent polymorphisms were present in the original wild populations.

Acknowledgments.—We should like to thank Professor J. L. Jinks for helpful discussions and the School of Education, University of Birmingham, for a programme for factor rotation. Computations were carried out on the K.D.F.9 computer at the University of Birmingham. This work was supported by the Agricultural Research Council and Medical Research Council.

7. References

BAILEY, N. T. J. 1961. *Introduction to the Mathematical Theory of Genetic Linkage.* Oxford University Press.

CLARKE, M. R. B. 1970. A rapidly convergent method for maximum-likelihood factor analysis. *Brit. J. Math. & Statist. Psychol., 23,* 43-52.

FALCONER, D. S. 1960. *Introduction to Quantitative Genetics.* Oliver and Boyd, Edinburgh.

GALE, J. S., AND ARTHUR, A. E. 1972. Variation in wild populations of *Papaver dubium.* IV. A survey of variation. *Heredity, 28,* 91-100.

JÖRESKOG, K. G. 1963. *Statistical Estimation in Factor Analysis.* Almqvist and Wiksell, Stockholm.

JÖRESKOG, K. G. 1967. Some contributions to maximum likelihood factor analysis. *Psychometrika, 32,* 443-482.

JÖRESKOG, K. G., AND LAWLEY, D. N. 1968. New methods in maximum likelihood factor analysis. *Brit. J. Math. & Statist. Psychol., 21,* 85-96.

KAISER, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187-200.

LAWLEY, D. N. 1940. The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edinb., A, 60,* 64-82.

LAWLEY, D. N., AND MAXWELL, A. E. 1971. *Factor Analysis as a Statistical Method,* 2nd edition. Butterworth, London.

LAWRENCE, M. J. 1972. Variation in wild populations of *Papaver dubium.* III. The genetics of stigmatic ray number, height and capsule number. *Heredity, 28,* 71-90.

MATHER, K. 1951. *The Measurement of Linkage in Heredity,* 2nd edition. Methuen, London.

MODE, C. J., AND ROBINSON, H. F. 1959. Pleiotropism and the genetic variance and covariance. *Biometrics, 15,* 518-537.

MORRISON, D. F. 1967. *Multivariate Statistical Methods.* McGraw-Hill, New York.

THURSTONE, L. L. 1945. *Multiple Factor Analysis.* University of Chicago Press, Chicago.